On Fixed Point error analysis of FFT algorithm

Shaik Qadeer[1], Mohammed Zafar Ali Khan[2], and Syed Abdul Sattar[3]
[1] MJCET/ Electrical, Hyderabad, India
Email: haqbei@gmail.com
[2] IITH/Electrical, Hyderabad, India
Email: zafar @iith.ac.in
[3] RITS/Electrical, Hyderabad, India
Email: syedabdulsattar1965@ gmail.com

***Abstract*—In this correspondence the analysis of overall quantization loss for the Fast Fourier Transform (FFT) algorithms is extended to the case where the twiddle factor word length is different from the register word length. First, a statistical noise model to predict the Quantization error after the multiplication of two quantized signals, of different precision, is presented. This model is then applied to FFT algorithms. Simulation results, that corroborate the theoretical analysis, are then presented.**

***Index Terms*— DFT (Discrete Fourier Transform), FFT (Fast Fourier Transform), DIT (Decimation in Time), and Quantization loss analysis.**

I. Introduction

The discrete Fourier transforms (DFT) and linear filtering is among the most fundamental operations in digital signal processing. The Fast Fourier transform is an algorithm to efficiently compute the discrete Fourier transform (DFT). It is a very useful algorithm, playing an important role in various digital signal processing applications from telecommunication, image processing, radar, sonar to vibrational analysis and material analysis and etc. In the actual hardware design, the accuracy of FFT/IFFT module is an important design factor of system performance. When it is implemented on a digital machine, quantization errors will arise due to the finite word length of the machine. Theoretical performance evaluation of signal to quantization noise (SQNR) of different FFT algorithms has been widely reported in previous works, for example [2]-[21]. All this consider the twiddle factor bit width to be similar to register bit width, as it simplifies the analysis. However, in many practical cases like fixed point DSP processors [1], the input bit widths are not the same, and the theoretical analysis given by [2]-[21] do not predict the saturation of the SQNR curve due to the constant twiddle factor bit width.

In this paper a model of quantization noise for multiplication when the input registers have different bit widths is developed first. The output noise of such a multiplier is then computed. The results are then applied to FFT algorithms and simulation results are presented to verify the accuracy of the proposed model

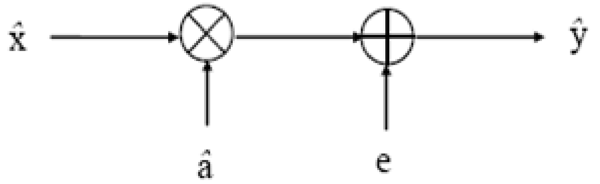

Figure1. Additive noise model of quantization loss for same bit width multiplication.

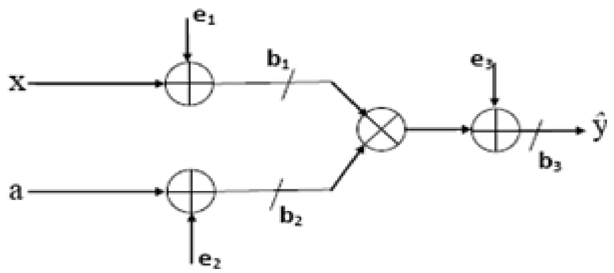

Figure2. Additive noise model of quantization loss for unequal size bit width multiplication.

The organization of paper is as follows: The multiplier quantization noise model that takes care of effects of different bit widths at the inputs is discussed in section II. Application to FFT algorithms is discussed in section III. Section IV gives Simulation results followed by Summary in section V.

II. GENERALIZED MULTIPLIER MODEL

The additive noise model of quantization loss is widely adopted to measure the effect of the fixed length operations in digital signal processing systems [2], [4]. The quantized product can be expressed as the sum of an unquantized product and a uniformly distributed additive quantization noise.

A. Previous model

If we consider the multiplication of quantized numbers $\hat{x}$ and $\hat{a}$ of bit width b, the product $\hat{y}$ is quantized to (b+1) bits, so that $\hat{y} = Q_b[y]$. The variance of this is given in [6] and the corresponding model is shown in Figure 1.

B. Proposed model

Consider the multiplication of quantized numbers $\hat{x}$ and $\hat{a}$ of a bit widths *b1* and *b2* respectively. The product $\hat{y}$ is

This work is a part of PhD thesis of Shaik Qadeer

DOI: 01.IJIT.01.03.508

quantized to b3 bits, so that $\hat{y} = Q_{b3}[y]$.

Each quantized number a, quantized to bit width *b*, can be represented as an unquantized number with an additive quantization noise source *e* [2], [4] as

$$\hat{a} = a + e \tag{1}$$

Where *e* is a uniformly distributed random variable whose probability density function (pdf) is given in equation (2), and variance is given by $\sigma^2 = \Delta^2/_{12}$ where

$\Delta = 2^{-b}$.

$$p(e) = \begin{cases} \frac{1}{\Delta}, \frac{-\Delta}{2} \le e \le \frac{\Delta}{2} \\ 0 \quad otherwise \end{cases} \tag{2}$$

The quantized product term can be expressed as the product of quantized inputs with an additive quantization noise source, *e3*, as

$$\hat{y} = Q_{b3}[y] = \hat{x}\hat{a} + e3 \tag{3}$$

If $b1 \neq b2 \neq b3$ then using equation (1) to replace quantized $\hat{x}$, $\hat{a}$ by their unquantized values we get

$$\hat{y} = (x + e1)(a + e2) + e3 = xa + n \tag{4}$$

where *n = e1a+e2x+e1e2+e3* is the noise term. The conditional variance of *n* given *x*, *a* is

$$\sigma_n^2 = \frac{2^{-2b3}}{12} + x^2\frac{2^{-2b2}}{12} + a^2\frac{2^{-2b1}}{12} + \frac{2^{-2(b1+b2)}}{12} \tag{5}$$

For the special case when *b1 = b2 = b3 = b*, we have

$$\sigma_n^2 = \frac{2^{-2b}}{12}\{1 + x^2 + a^2 + 2^{-2b}\} \tag{6}$$

Denoting $E\{|x|^2\} = \sigma_x{}^2, E\{|a|^2\} = \sigma_a{}^2$, the variance of n for different bit width input is given by

$$\sigma_n^2 = \frac{1}{12}\{2^{-2b3} + \sigma_x^2 2^{-2b2} + \sigma_a^2 2^{-2b1} + 2^{-2(b1+b2)}\} \tag{7}$$

Assuming as in [[2], equation 6.4.7], that *x* and *a* are uncorrelated and have uniform density in the range $(^{-1}/_N, ^1/_N)$ we have

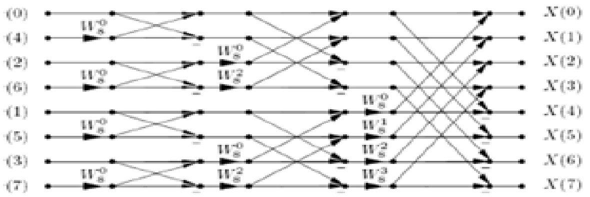

Figure 3-Flow-graph of the DIT- FFT algorithm.

$$\sigma_x{}^2 = \sigma_a{}^2 = \frac{1}{3N^2} \tag{8}$$

Substituting equation (8) in equation (7), we obtain

$$\sigma_n^2 = \frac{1}{12}\{2^{-2b3} + \frac{1}{3N^2}[2^{-2b2} + 2^{-2b1}] + 2^{-2(b1+b2)}\} \tag{9}$$

Figure 2 depicts quantized product term $\hat{y}$ as the product of quantized inputs with an additive quantization noise source having different bit widths.

III. APPLICATION TO FFT

In this section the generalized multiplier model developed in previous section is applied to FFT. For this case we assume, without loss of generality, that *x* is an input to the FFT, and *a* is the twiddle factor. Then variance of *a* will be $\sigma_a{}^2 = 1$ and assuming as in [[2], eqn. 6.4.7], that *x* and *a* are uncorrelated and *x* has uniform density in the range $(^{-1}/_N, ^1/_N)$, the variance of noise for each multiplication, given in equation (7), specializes to

$$\sigma_n^2 = \frac{1}{12}\{2^{-2b3} + \sigma_x^2 2^{-2b2} + 2^{-2b1} + 2^{-2(b1+b2)}\} \tag{10}$$

A. Error analysis of Radix-2 FFT algorithm

In this subsection we discuss the variance of QE for Radix-2 FFT algorithm to the case of different register bit width. From the flow graph of the DIT FFT algorithm given in Figure 3a, it can be seen that the DFT samples are computed by a series of butterfly computations with a single complex multiplication per butterfly module. Some of the butterfly computations require multiplications by *- 1* or *-j* that we do not treat separately here, to simplify the analysis. From Figure 3a, it is also observed that in general there are *N/2* complex multiplications in first stage, *N/4* in the second stage, *N/8* in third stage, and so on, until the last stage, where there is only one complex multiplication. Following the procedure as in [2], instead of scaling the input samples by *1/N*, we can distribute the total scaling of 1/N into each of the FFT stages to avoid overflow i.e. we can scale the input signals at each stage by 1/2. This scaling reduces the variance of QE as follows. Each factor of 1/2 reduces the variance of QE by a factor of 1/4. Thus 4(N/2) QE introduced in first stage will reduced the variance by $(1/4)^{v-1}$, the 4(N/4) in second stage to $(1/4)^{v-2}$, and so on, where is the number of FFT stages. Hence, the total variance of the QE at the output of FFT algorithm will be

$$\sigma_q^2 = \sigma_n^2\{4(\frac{N}{2})(\frac{1}{4})^{\nu-1} + 4(\frac{N}{4})(\frac{1}{4})^{\nu-2} + ... + 4\} \tag{11}$$

which simplifies as [2], we get

$$\sigma_q^2 = 8\sigma_n^2\{1 - (\frac{1}{2})^{\nu}\} \tag{12}$$

© 2011 ACEEE
DOI: 01.IJIT.01.03.508

For large values of *N*, FFT size, this can be approximated as

$$\sigma_q^2 \approx 8\sigma_n^2 \qquad (13)$$

Due to the scaling the input, the variance of the signal at the output of FFT will become $\sigma_X{}^2 = {}^1\!/_{3N}$ and SQNR is given by

$$\frac{\sigma_X{}^2}{\sigma_q{}^2} = \frac{1}{24N\sigma_n{}^2} \qquad (14)$$

B. Error analysis of Split radix DIT FFT algorithm

In this subsection we consider QE for Split radix FFT algorithm. From the block diagram as shown in Figure 3b, it is clear that each butterfly computation invloves 2 complex or 8 real multiplications. The number of butterflies from stage k=2 to is given by

$$B_k = [2 + (-\frac{1}{2})^{\nu-k}]\frac{N}{12} \qquad (15)$$

and the number of Radix-2 butterflies in stage *k=1* is given by

$$B_{R2} = [\frac{2^{\nu} - (-1)^{\nu}}{3}] \qquad (16)$$

As Radix-2 multiplications are all non-trivials, so need not to be consider for QE analysis. Now the variance of the QE for this case for the computation of *N*-point DFT is given as

$$\sigma^2_{q/N} = \sigma_n^2 8\{[2 + (-\frac{1}{2})^{\nu-2}]\frac{N}{12}(\frac{1}{4})^{\nu-3} + \\ \ldots + [2 + (-\frac{1}{2})^{\nu-\nu}]\frac{N}{12}(\frac{1}{4})^{\nu-(\nu+1)}\} \qquad (17)$$

which simplifies as [2], we get

$$\sigma^2_{q/N} = \frac{64N}{9}\sigma_n^2\{[1-(\frac{1}{4})^{\nu-1}] + {}^1\!/_3[1-(-\frac{1}{8})^{\nu-1}]\} \qquad (18)$$

For large values of *N*, FFT size, this can be approximated as

$$\sigma^2_{q/N} \approx 9.48N\sigma_n^2 \qquad (19)$$

QE for the computation of particular split radix FFT output is

$$\sigma^2_{q/1} \approx 9.48\sigma_n^2 \qquad (20)$$

Equation (20) is the noise variance of split radix DIT FFT algorithm due to quantization . SQNR for this case is given by

$$\frac{\sigma_X{}^2}{\sigma^2_{q/1}} = \frac{1}{28.44N\sigma_n{}^2} \qquad (21)$$

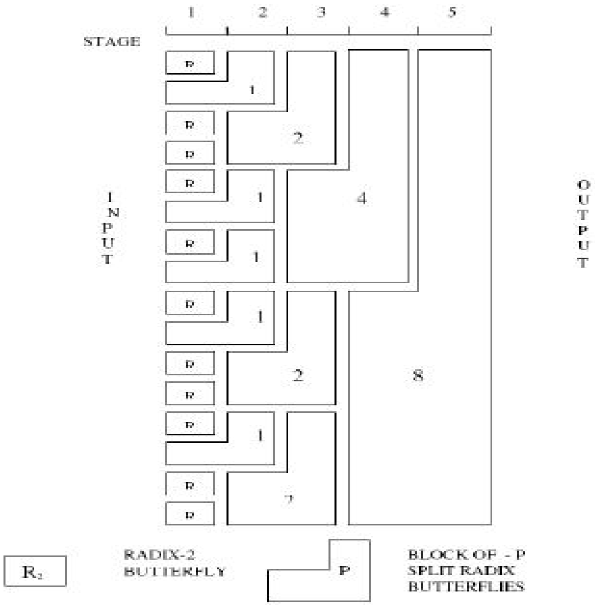

Figure 3b-Flow-graph of the SRDIT- FFT algorithm for N=32.

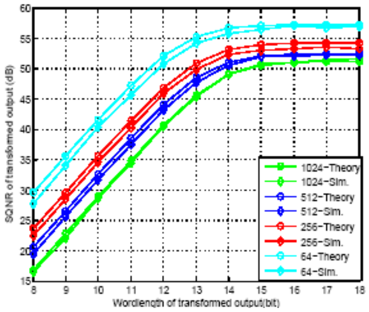

Figure 4. SQNR comparison chart of Radix-2DIT FFT algorithm with fixed twiddle factor (10bits).

IV. COMPARATIVE SIMULATION RESULTS

In order to verify the expression derived in the previous section, a fixed point simulation of SQNR for different FFT size is presented. It is assume that the word length of the internal register is same as that of the output register (b1 = b3). Figure 4 shows SQNR of Radix-2 DIT FFT algorithms with the word length of twiddle factor set to 10 bits (b2 = 10), and the internal word length of fixedpoint FFT is swept from 8 to 18 bits. FFTs of length 64, 256, 512 and 1024 are simulated. From the figure it can be observed that the simulated SQNR (denoted by 'NSim.' for N-point FFT in Fig. 4) is within 0.5 dB of theoretical SQNR (denoted by 'N- Theory'). Similar simulation for split radix FFT is shown in figure 5. For cross verification figure 6 plots the SQNR as a function of FFT length for various values of twiddle factor and internal register word lengths. It is observed that the theoretical SQNR is within 0.5 dB of the simulated SQNR. Accordingly, it can be concluded that the simulation results closely match theoretical SQNR curves obtained by plotting equation (14).

© 2011 ACEEE
DOI: 01.IJIT.01.03.508

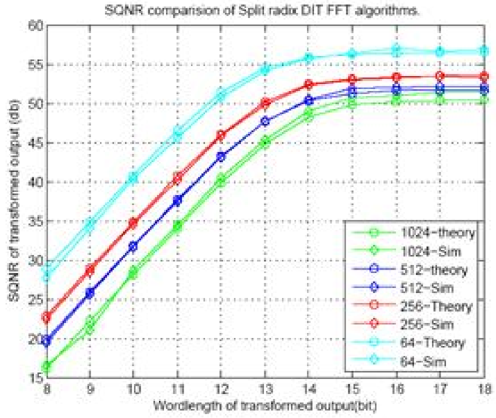

Figure 5. SQNR comparison chart of Split radix DIT FFT algorithm with fixed twiddle factor (10bits).

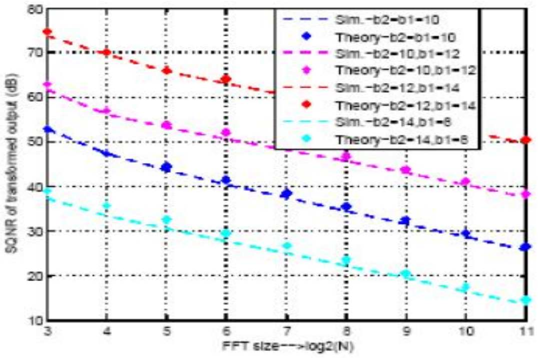

Figure 6. SQNR comparision of Radix-2DIT FFT algorithms with fixed input (10,12 bits).

V. CONCLUSION

We have developed a generalized multiplier model and applied it to derive the signal to quantization noise for FFT, for the case when twiddle factor word length is different from register word length. The results obtained are an important basis for the implementation of the FFT algorithm. Simulation results indicate that the theoretical analysis agrees closely with the actual behavior of SQNR.

ACKNOWLEDGMENT

The authors wish to thank MJCET for allowing us to conduct this research. This work was supported in part by a partial grant from MJCET.

REFERENCES

[1] R. Venketramani and M. Bhaskar, "*Digital Signal Processors*", Tata McGraW Hill, India, 2007.

[2] John G. Proakis and D.G. Manolakis, "*Digital Signal Processing-Principles, Algorithms, and Applications*", 3rd edition, Chapter 6, Prentice Hall of India, 2003.

[3] R.B. Perlow and T.C. Denk, "Finite Wordlength Design for VLSI FFT Processors," *in Conf. Rec. 35th Asilomar Conf. Signals, Systems, Computers*, 2001, vol. 2-2, pp. 1227-1231

[4] W.-H. Chang and T. Nguyen, "On the Fixed-Point Accuracy Analysis of FFT Algorithms," *IEEE Transactions On Signal Processing*, Vol. 56, No. 10, pp. 4673-4682, October 2008.

[5] T. Tran, B. Liu, "Fixed-point fast Fourier transform error analysis", *IEEE Trans. on ASSP*, 1976, vol.24(6), pp. 563–573.

[6] W.-H. Chang and T. Nguyen, "Integer FFT with optimized coefficient sets," *in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*,pg. 109-112, 2007.

[7] T. Kaneko and B. Liu, "Accumulation of roundoff errors in fast Fourier transforms", *J. Ass. Comput.* Mach., vol. 17, pp. 537–654, Oct. 1970.

[8] C. J. Weinstein, "Roundoff noise in floating point fast Fourier transform computation", *IEEE Trans. Audio Electroacoust.*, vol. AU-17,pp. 209–215, Sept. 1969.

[9] G. U. Ramos, "Roundoff error analysis of the fast Fourier transforrn", *Math. Comput.*, vol. 25, pp. 757-768, Oct. 1971.

[10] S.Y. Park and N.I. Cho, "Fixed point error analysis of CORDIC processor based on the variance propagation formula", *IEEE Trans. Circuits, Sys.I,* Reg. Papers,vol. 51,no. 3, pp. 573-584, 2004.

[11] Wade Lowdermilk and Fred Harris, "Finite Arithmetic Considerations for the FFT Implemented in FPGA-Based Embedded Processors in Synthetic Instruments", *in IEEE Instrumentation and Measurement Magazine*, vol.8, no. 3, pg. 40-46, August 2007.

[12] P.D. Welch, "A fixed point fast Fourier transform error analysis", *in IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 151–157, June 1969.

[13] D.V. James, "Quantization errors in fast Fourier transform", *in IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-23, pp. 277–283, June 1975.

[14] A. V. Oppenheim and R. Schafer, "*Digital Signal Processing*", Chapter-9, Prentice-Hall of India, 2002.

[15] Sanjit Mitra, "*Digital Signal Processing: A Computer Approach*", 3rd edition, Chapter-12, McGraw-Hill, 2006.

[16] C.W. Barnes,B.N. Tran and S.H. Lueng, "On the Statistics of Fixed-Point roundoff error," *in IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 3,June-1985.

[17] V. Ashok N. and K.N.N. Prabhu,"The fractional Fourier transform: theory, implementation and error analysis", *in Elsevier Journal on Microprocessors and Microsystems*, vol. 27 , pp.511-521, 2003.

[18] D. Chandra, "Accumulation of coefficient roundoff error in FFT implemented with logarithmic number system", *in IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 11, pp. 1633-1636, 1987.

[19] W. Schlecker, Christiane B. and H. Pfleiderer, "Quantisation Noise in Fixed-Point Multiplications," *in Electrical Engineering (Archiv fur Elektrotechnik)* Volume 89, Number 4, 339-342, DOI: 10.1007/s00202-006-0009-3.

[20] A. S. Sripad and D. L. Snyder,"A necessary and sufficient condition for quantization errors to be uniform and white," *in IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 442-448, Oct. 1977.

[21] Z. Lukac and M. Temerinac , "Analysis of some methods For maintaining accuracy in implementation of FFT on fixed point DSP ," *in IEEE conerence, Serbia and Montenegro*, Nis, September 28 -30, 2005.

A part of paper is presented in CEMC2011

DOI: 01.IJIT.01.03.508